THE POTENTIAL OF THE IPI PAN CORPUS*

Adam Przepiórkowski
Institute of Computer Science, Polish Academy of Sciences, Warsaw

ABSTRACT

The aim of this article is to present the IPI PAN Corpus (cf. `http://korpus.pl/`), a large morphosyntactically annotated XML encoded corpus of Polish developed at the Institute of Computer Science, Polish Academy of Sciences, and describe its potential and actual use in natural language processing and in linguistic research. In particular, various quantitative information about the corpus and its publicly available subcorpora is provided including: sizes in terms of orthographic words and interpretable segments, tagset size measured in types and tokens, etc., but also information reflecting interesting facts about Polish, i.e., frequencies of words of different lengths, frequencies of grammatical classes and some grammatical categories, and initial frequency lists of Polish lemmata.

1. Introduction

The IPI PAN Corpus, a large morphosyntactically annotated XML[1] encoded corpus of Polish, is one of the results of a corpus project financed by the State Committee for Scientific Research (Polish: Komitet Badań Naukowych; project number 7 T11C 043 20) from mid-2001 to mid-2004, henceforth, the IPI PAN Corpus project, carried out at the Institute of Computer Science, Polish Academy of Sciences (Polish: Instytut Podstaw Informatyki PAN; hence, the IPI PAN Corpus). Other results of the project, documented elsewhere, include: the design of a flexemic tagset for Polish (Woliński 2003, Przepiórkowski 2003, Przepiórkowski and Woliński 2003), a statistical tagger (Dębowski 2003, 2004) and the Poliqarp search engine featuring an expressive query language (Przepiórkowski et al. 2004, Przepiórkowski 2004).

The aim of this article is to quantitatively characterize the corpus[2] and describe its potential in natural language processing (NLP) and in linguistic

* The author is grateful to Łukasz Dębowski, Jakub Fast, Elżbieta Hajnicz, Anna Kupść, Agnieszka Mykowiecka, Maciej Piasecki, Tadeusz Piotrowski, Zygmunt Saloni, Agata Savary, Beata Trawiński and Marcin Woliński for comments on earlier versions of this paper, and to Przemysław Kaszubski for the invitation to the workshop on *Assessing the potential of corpora* and for his patience.

[1] More precisely: XCES, i.e., XML Corpus Encoding Standard; cf. Ide et al. 2000.

[2] All quantitative information given below pertains to the so-called 'preliminary' version of the IPI PAN Corpus (Przepiórkowski 2004) of June 2004.

research.

The paper starts with a brief presentation of the approach to text segmentation and morphosyntactic tagging adopted in the IPI PAN Corpus, §2. The following section, §3, provides information about the size of corpus measured in segments and in orthographic words and about the size of the tagset, as well as various frequency information reflecting interesting facts about Polish, such as frequencies of words conditioned on their lengths, frequencies of grammatical classes, frequencies of some grammatical categories, and a short list of the most frequent Polish lemmata. The next section, §4, shows the potential of the IPI PAN Corpus for discovering linguistic facts which go beyond words out of context (identification of bound words, automatic acquisition of subcategorisation frames, support in the development of hierarchical WordNet-like thesauri) and for certain NLP tasks (morphosyntactic disambiguation). The article concludes with §5, which also mentions the usefulness of the IPI PAN Corpus in education.

2. Segmentation and tagset

In order to appreciate the quantitative results presented below, it is necessary to understand basic assumptions about the segmentation (tokenisation) procedure and the tagset employed in the IPI PAN Corpus. Both are described in detail in Woliński 2003, Przepiórkowski 2003, and Przepiórkowski and Woliński 2003, with the final version of the tagset presented in Przepiórkowski 2004. The present section offers a concise summary of the most important aspects of segmentation and tagging.

2.1. Segmentation

Segments are those sequences of orthographic characters to which tags are assigned. Segments in this sense are often understood as orthographic words (approximately, words "from space to space") and, in fact, segments of the IPI PAN Corpus are never longer than such orthographic words. However, in some special cases, segments may be shorter than orthographic words and, moreover, some non-words sequences of characters, especially punctuation marks, are assigned tags, i.e., they are treated as segments.

Cases where orthographic words are split into smaller segments include first and second person past forms such as *łgał*|*eś* 'lied-you', *długo*|*śmy* 'long time-we', *tak*|*em* 'so-I', etc., words containing particles such as *by* (subjunctive particle), *-ż(e)* (emphatic particle) and *-li* (question particle), e.g., *przyszedł*|*by* 'come-would', *napisała*|*by*|*m* 'write-would-I', *chodź*|*że* 'come-Emph', *potrzebował*|*że*|*by*|*ś* 'need-Emph-would-you', *znasz*|*li* 'know-

Q', prepositions incorporating post-prepositional weak pronominal form *-ń*,[3] as in *do*|*ń* 'to-him' or *ze*|*ń* 'with-him', and also some words containing the hyphen, i.e., words such as *polsko*|-|*niemiecki* 'Polish-German' and double names, e.g., *Kowalska*|-|*Nowakowska*.

2.2. Tagset

In the IPI PAN Tagset, each morphosyntactic tag is a sequence of colon-separated values, e.g.: subst:sg:nom:m1 for the segment *chłopiec* 'boy'. The first value, e.g., subst, determines the *grammatical class*, i.e., roughly, part of speech (POS), while the values that follow it, e.g., sg, nom and m1, are the values of grammatical categories appropriate for that grammatical class.

Grammatical categories adopted here are: *number* (sg, pl), *case* (nom, gen, dat, acc, inst, loc, voc), *gender* (m1, m2, m3, f, n), *person* (pri, sec, ter), *degree* (pos, comp, sup), *aspect* (imperf, perf), *negation* (aff, neg), *accentability* (akc, nakc), *post-prepositionality* (praep, npraep), *accommodability* (congr, rec), *agglutination* (nagl, agl) and *vocalicity* (wok, nwok).

Grammatical classes are based on the notion of *flexeme* introduced in Bień 1991, 2004 — they are inflectionally uniform subsets of more traditional lexemic classes (POSs). The following grammatical classes are used in the IPI PAN Corpus: nominal classes: noun (subst; e.g., *profesorowie*), depreciative form (depr; e.g., *profesory*); pronominal classes: non-3rd person pronoun (ppron12; e.g., *nas*), 3rd-person pronoun (ppron3; e.g., *oni*), strong reflexive pronoun SIEBIE (siebie; e.g., *sobą*); numeral (num; e.g., *pięciu*); adjectival classes: adjective (adj; e.g., *polski*), ad-adjectival adjective (adja; e.g., *polsko*), post-prepositional adjective (adjp; e.g., *polsku*); adverb (adv; e.g., *długo*); (de-)verbal classes: non-past form (fin; e.g., *zrobi*), future BYĆ (bedzie; e.g., *będą*), agglutinate BYĆ (also called *mobile inflection*; aglt; e.g., *-śmy*), l-participle (praet; e.g., *zrobił*), imperative (impt; e.g., *zrób*), impersonal (imps; e.g., *zrobiono*), infinitive (inf; e.g., *zrobić*), contemporary adverbial participle (pcon; e.g., *robiąc*), anterior adverbial participle (pant; e.g., *zrobiwszy*), gerund (ger; e.g., *zrobienie*), active adjectival participle (pact; e.g., *robiący*), passive adjectival participle (ppas; e.g., *zrobiony*), winien (winien; e.g., *powinien*), predicative (pred; e.g., *szkoda*, *widać*, *to*); functional classes: preposition (prep; e.g., *na*), conjunction (conj; e.g., *oraz*), particle-adverb (qub; e.g., *by*, *nie*); other classes: nominal alien (xxs), other alien (xxx), unknown form (ign) and punctuation (interp).

[3] In the automatically annotated parts of the IPI PAN Corpus, such contractions are often not recognised as consisting of multiple segments.

3. Words

This section presents some hitherto unpublished quantitative information about the IPI PAN Corpus and, especially, about two publicly available subcorpora: the smaller (around 15 million segments; see below) and more balanced `sample` corpus, searchable via the Internet at `http://korpus.pl/`, and a larger (around 70 million segments) `wstepny` corpus downloadable for searching from the same site. Similar data about `frek`, the older tiny (0.5 million running words) corpus of the "Frequency dictionary of contemporary Polish" (Kurcz et al. 1990), compiled in late 1960ies and early 1970ies (Kurcz et al. 1974), subsequently cleaned-up (cf. Ogrodniczuk 2003 and references therein) and manually re-tagged within the IPI PAN Corpus project, are also cited.

3.1. Segments

The IPI PAN Corpus as a whole is heavily unbalanced: most of the text in the corpus comes from newspapers, transcripts of parliamentary sessions and legal texts. Also `wstepny` consists mainly of parliamentary proceedings (over 68%) and newspapers (almost 21%), with only 6.5% of artistic prose, 3% of scientific texts and 1% of legal texts.

Some effort towards so-called representativeness was put into the make-up of the `sample` corpus, which consists of scientific texts (10%), contemporary artistic prose (10.6%), older (late XIX and early XX century) artistic prose of the kind read at schools (9.7%), legal texts (4.9%), transcripts of parliamentary sessions (15.5%) and various newspaper texts (49.3%). As is well known, also the `frek` corpus of the "Frequency dictionary of contemporary Polish" is supposed to be balanced, with 20% of popular science, 20% of news dispatches, 20% of editorials and longer articles, 20% of artistic prose, and 20% of artistic drama.

The sizes of the IPI PAN Corpus as a whole, the two subcorpora `sample` and `wstepny`, as well as the `frek` corpus, are given in Table 1. The **segments** column contains the exact number of segments (including punctuation), while the **words** column shows the exact number of orthographic words (excluding punctuation) in each of these corpora. The final column gives the segment-to-word ratio calculated on the basis of the previous two columns.[4]

[4] Note that the number of words for `frek` in Table 1 differs from the declared 500,000. The difference probably stems from the fact that, as noted in Kurcz et al. 1990 (see also Czerepowicka and Saloni 2004), in the original version of `frek`, multiple orthographic words were sometimes treated as single wordforms containing a space, e.g., the reflexive marker *się* following an adjectival participle or a gerundial form, foreign surnames containing *de*, *von*, etc., fossilised sequences of prepositions and bound words (see §4.2 below), etc.

	segments	words	ratio
IPI PAN Corpus	360,446,336	291,187,457	1.24
`wstepny`	70,492,786	58,317,809	1.21
`sample`	15,252,022	12,198,241	1.25
`frek`	659,511	545,970	1.21

Table 1: Corpus sizes measured in segments and orthographic words.

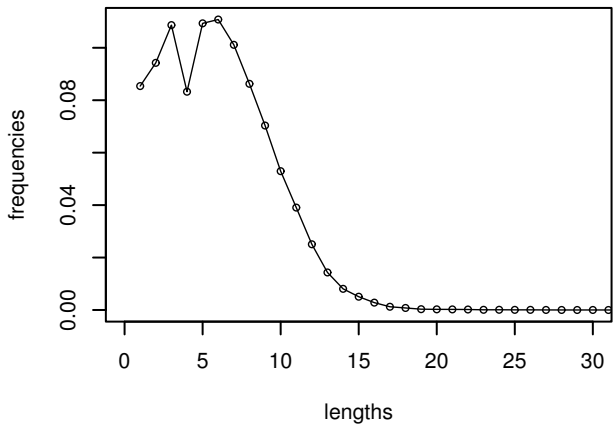

Figure 1: Frequencies of lengths of words in `frek`.

Additionally, Fig. 1–2 show the frequencies of words of various lengths in `frek` and `sample`. It is clear from these figures that the most frequent number of letters in words is 6 (`frek`, as well as `wstepny` and the whole IPI PAN Corpus; the last two not illustrated here) or perhaps 5 (`sample`), and that, interestingly, words of length 4 are conspicuously less frequent than words of neighbouring lengths.

Finally, let us mention that the mean length of words calculated on the basis of these corpora are: 5.92 (`frek`), 5.78 (`sample`), 5.94 (`wstepny`) and 5.90 (the whole IPI PAN Corpus).

3.2. Tags

As mentioned above, each complete tag is a list whose first element is a grammatical class and the other elements are values of grammatical categories appropriate for this class. For example, each nominal tag has the form subst:*number*:*case*:*gender*, where *number* is sg or pl, *case* is one of 7 cases and *gender* is one of 5 genders, i.e., there are 70 potential nominal tags. Overall, the current tagset allows for 4179 potential tags, although many of the potentially possible combinations of grammatical classes and

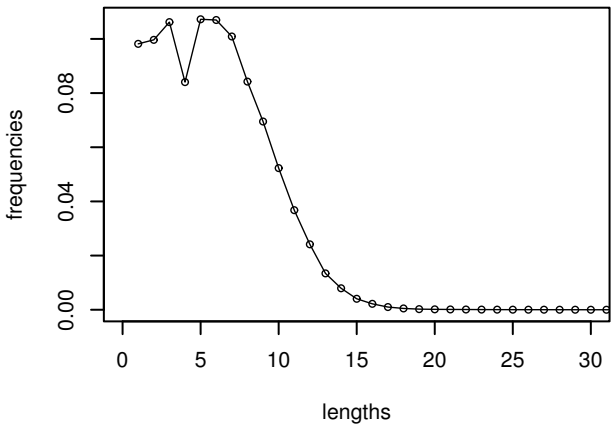

Figure 2: Frequencies of lengths of words in `sample`.

grammatical categories are never realised. The number of different tags actually found in corpora is given in Table 2.

	known segments		**unknown segments**		**all segments**	
	disamb.	**all**	**disamb.**	**all**	**disamb.**	**all**
`wstepny`	945	1149	259	368	946	1150
`sample`	912	1131	237	357	913	1132
`frek`	–	–	–	–	1642	1642

Table 2: Actual tags — only those disambiguated or all tags proposed by the morphological analyser Morfeusz — for segments known by Morfeusz, for unknown segments, whose morphological interpretation was guessed on the basis of Tokarski 1993, and for all segments.

The most conspicuous difference between the rows for `frek` on the one hand and for `sample` and `wstepny` on the other hand stems from the fact that `frek` is annotated manually, with tags which are correct in a given context (we will call them 'disambiguated tags'), hence, there is no difference in `frek` between disambiguated tags and all tags. This also implies that, in `frek`, there are no segments marked as unknown to the morphological analyser, hence the '–'s.

On the other hand, as explained in more detail in Przepiórkowski 2004, not only does the IPI PAN Corpus contain appropriately marked disambiguated tags, but it also retains all other tags proposed by the morphological analyser used in the project, i.e., by Morfeusz (by Zygmunt Saloni and Marcin Woliński). Moreover, since Morfeusz is a dictionary-based anal-

yser, there are words unknown to it.[5] In such cases, a guesser derived from Tokarski 1993 is used, which proposes interpretations on the basis of endings of words. As Table 2 shows, the repertoire of tags proposed by the guesser is quite limited with respect to the tags returned by Morfeusz (e.g., 368 vs. 1149 for `wstepny`); for example, the guesser never proposes pronominal, conjunctive or prepositional tags. Also, the only tag produced by the guesser which is not produced by Morfeusz is ign (hence, the difference of one between the **known** and **all** column pairs).

Another observation which begs explanation is that quite many tags returned by Morfeusz are never selected by the statistical disambiguator of Dębowski 2004: there are 1149 − 945 = 204 such tags in `wstepny` and 1131 − 912 = 219 in case of `sample`. In both cases the source of discrepancy is roughly the same; for example, in the latter case, 152 of those tags are participial tags (almost all adjectival participles), 36 — pronominal tags mainly with genders m2 and m3, 20 — comparative and superlative adjectival tags, 7 — vocative numeral and nominal tags, and a few odd prepositional and gerundial tags. It may be hypothesised that this difference reflects some systematic errors made by the disambiguator.[6]

Finally, the difference between the tags present in `frek` (1642) and all the tags present in, e.g., `sample` (1132) is caused by 930 tags present in `frek` but not in `sample`, and 420 tags present in `sample`, but not in `frek`. The majority of the 930 tags found in `frek` only are manual annotation errors resulting in such inconsistent tags as ppron3:pl:acc:f:pri, i.e., a 3rd person pronoun in plural number, accusative case, feminine gender and... first person. Three other main classes of `frek`-only tags are: numerals with explicit congr/rec information (65; automatic annotation often left those underspecified), vocative forms (50) probably not known to Morfeusz and various alien tags (automatic tagging was not able to distinguish alien forms from other unknown forms). On the other hand, surprisingly, the 420 `sample`-only tags are mainly the 264 adjectival participial (ppas and pact) tags and 130 first and second pronoun (ppron12) tags.

The results reported above show that much work still needs to be carried out to alleviate the annotation problems introduced both by manual and by automatic annotation.

Let us finish by comparing the frequencies of the main groups of grammatical classes, of case values and of gender values in `frek` and `sample`. In Fig. 3–4, the group *noun* comprises classes subst and depr, *verb* — all (de-)verbal classes enumerated in §2.2, *adjective* — the three adjectival classes mentioned there, *pronoun* — the pronominal classes ppron12, ppron3

[5] As reported in Piskorski et al. 2004, Morfeusz does not recognise about 5% of wordforms.
[6] Dębowski 2004 reports the 9.4% error rate of the disambiguator.

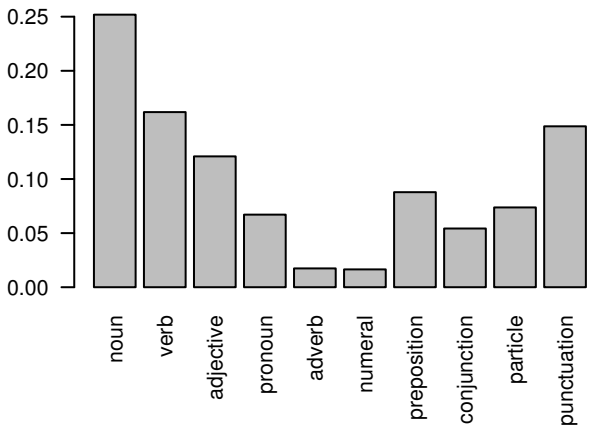

Figure 3: Frequencies of grammatical classes in `frek`.

and siebie, while *adverb*, *numeral*, *preposition*, *conjunction*, *particle* and *punctuation* are adv, num, prep, conj, qub and interp, respectively. Since the breakdown presented in these figures is made on the basis of disambiguated tags of known forms, classes xxs, xxx and ign are not included. Two clear differences between these two corpora concern pronouns and numerals, both more frequent in `frek`. The difference in the frequencies of numerals might be caused by the fact that numbers are converted to words in `frek` and treated as numerals,[7] while there is no such conversion in the IPI PAN Corpus, which results in the assignment of ign. On the other hand, the differences in the frequencies of pronouns might be the result of the high percentage of pronoun-rich 'artistic drama' in `frek` (although, on the other hand, the IPI PAN Corpus contains a high percentage of parliamentary proceedings, which also to some extent approximates spoken language).

Finally, Fig. 5–8 compare the frequencies of cases and genders in `frek` and `sample`. As can be seen from these figures, `sample` contains a conspicuously lower percentage of dative forms and animate masculine forms. The question whether this difference reflects any real change in Polish between 1960ies (`frek`) and now (over 90% of texts in `sample` date from the last 5 years), or whether it simply results from the difference in genre breakdown and annotation procedure, is left for future research.

[7] Czerepowicka and Saloni 2004 note that this 'conversion' was actually an interpretative and non-deterministic process.

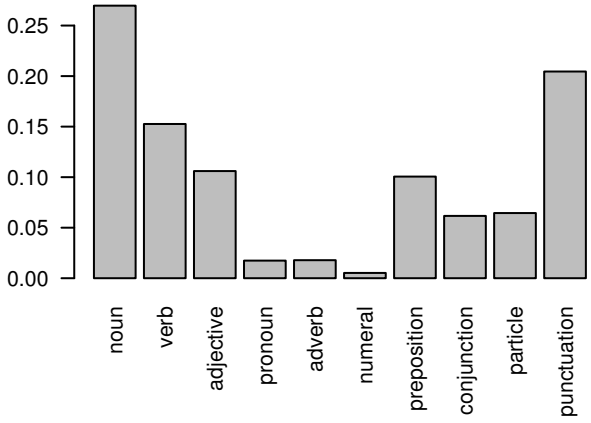

Figure 4: Frequencies of grammatical classes in `sample`.

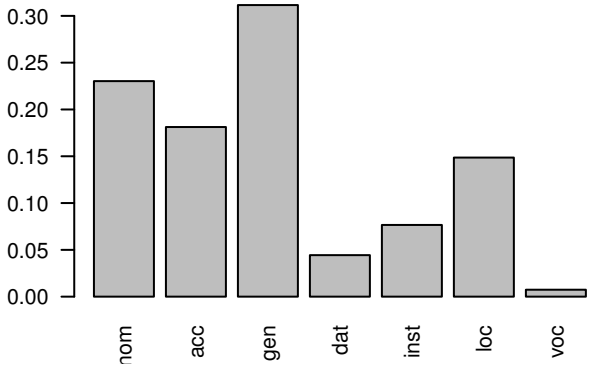

Figure 5: Frequencies of grammatical cases in `frek`.

3.3. Frequency lists

Given the various differences between `frek` and `sample`, it may be instructive to compare the so-called frequency lists, i.e., lists of the most frequent lexemes, ordered by frequency. Table 3 gives the 50 most frequent lexemes in both corpora, together with the exact count of their occurrences in these corpora and with their frequencies expressed as percents of all segments.

`frek`					`sample`			
W	prep	17501	2.65	1	W	prep	404573	2.65
I	conj	13304	2.02	2	I	conj	278221	1.82
ON	pron	11559	1.75	3	Z	prep	231889	1.52
SIĘ	qub	11104	1.68	4	BYĆ	verb	229609	1.51

BYĆ	verb	10860	1.65	5	SIĘ	qub	212322	1.39
NIE	qub	9765	1.48	6	NA	prep	202186	1.33
NA	prep	9729	1.48	7	NIE	qub	178164	1.17
Z	prep	9106	1.38	8	ON	pron	148075	0.97
DO	prep	6330	0.96	9	DO	prep	147492	0.97
TEN	adj	6159	0.93	10	TEN	adj	123217	0.81
ŻE	conj	4743	0.72	11	ŻE	conj	111352	0.73
TO	subst	3921	0.59	12	O	prep	96198	0.63
O	prep	3446	0.52	13	KTÓRY	adj	93557	0.61
A	conj	3258	0.49	14	A	conj	77742	0.51
KTÓRY	adj	3175	0.48	15	TO	subst	73188	0.48
MIEĆ	verb	2912	0.44	16	PAN	subst	60476	0.40
JAK	conj	2545	0.39	17	MIEĆ	verb	57947	0.38
CO	subst	2383	0.36	18	PO	prep	52675	0.35
ALE	conj	2219	0.34	19	OD	prep	44157	0.29
PO	prep	2146	0.33	20	CO	subst	43991	0.29
PAN	subst	2079	0.32	21	JA	pron	43972	0.29
OD	prep	2071	0.31	22	ALE	conj	43750	0.29
TAK	qub	1923	0.29	23	JAK	conj	42682	0.28
ROK	subst	1880	0.29	24	ROK	subst	41544	0.27
JUŻ	qub	1831	0.28	25	MÓC	verb	40241	0.26
MÓC	verb	1704	0.26	26	PRZEZ	prep	39741	0.26
TAKI	adj	1650	0.25	27	ZA	prep	37149	0.24
PRZEZ	prep	1557	0.24	28	BY	qub	33943	0.22
ZA	prep	1489	0.23	29	TAK	qub	32801	0.22
TO	pred	1470	0.22	30	DLA	prep	31763	0.21
DLA	prep	1459	0.22	31	JUŻ	qub	31724	0.21
BY	qub	1410	0.21	32	TO	pred	30583	0.20
TYSIĄC	subst	1374	0.21	33	BARDZO	adv	29767	0.20
JA	pron	1301	0.20	34	TAKI	adj	29507	0.19
SWÓJ	adj	1279	0.19	35	TYLKO	qub	26259	0.17
TYLKO	qub	1274	0.19	36	CZY	conj	26233	0.17
BARDZO	adv	1249	0.19	37	SWÓJ	adj	24359	0.16
JEDEN	adj	1239	0.19	38	ORAZ	conj	23055	0.15
CZŁOWIEK	subst	1201	0.18	39	POSEŁ	subst	22644	0.15
INNY	adj	1172	0.18	40	SPRAWA	subst	22139	0.15
WIEDZIEĆ	verb	1158	0.18	41	MÓWIĆ	verb	22055	0.14
SIEBIE	pron	1156	0.18	42	JEDEN	adj	22015	0.14
JESZCZE	qub	1140	0.17	43	ZOSTAĆ	verb	21501	0.14
PRZY	prep	1063	0.16	44	INNY	adj	21345	0.14
MÓWIĆ	verb	1054	0.16	45	SIĘ	pron	21039	0.14

PRACA	subst	1027	0.16	46	BO	conj	20907	0.14
PIERWSZY	adj	1019	0.15	47	TO	conj	20547	0.13
DWA	num	1015	0.15	48	PRZED	prep	20369	0.13
SAM	adj	1008	0.15	49	CHCIEĆ	verb	19695	0.13
CHCIEĆ	verb	997	0.15	50	DZIEŃ	subst	19530	0.13

Table 3: The most frequent lexemes in `frek` and `sample`

The comparison confirms that the preposition W 'in' is unquestionably the most frequent Polish lexeme, with the conjunction I 'and' as the second most frequent lexeme. Some differences in the two rankings may be due to chance or different genre make-up of the two corpora; in particular, the drastically different ranks of POSEŁ 'deputy, member of parliament' in `sample` (rank 39) and in `frek` (rank 1349) probably reflects the fact that transcripts of parliamentary sessions constitute around 15% of `sample`. On the other hand, the difference in the ranks of TYSIĄC 'thousand' probably stems from the fact that numbers were 'normalized' to words in `frek` (where TYSIĄC has rank 33) but not in the IPI PAN Corpus (rank 294 in `sample`); cf. Czerepowicka and Saloni 2004.

4. Beyond words

Counting forms occurring in corpora, and their lemmata and tags, is a rather simple, although already very useful use of natural language corpora, which only requires a simple collection of information that is already explicitly contained in a corpus. More interesting uses of linguistic corpora involve finding information which is only implicit in a corpus, which usually requires more or less advanced statistical reasoning. This section presents a few such tasks, all concerned with finding various relations between words.

4.1. Morphosyntactic disambiguation

The most obvious NLP use of manually disambiguated corpora, amply documented elsewhere, is for the construction of automatic disambiguators or taggers. The `frek` corpus, fully manually disambiguated within the IPI PAN Corpus project, as well as the small parts of the actual IPI PAN Corpus that were also manually disambiguated, have already been used for training automatic taggers, namely, in various versions of the tagger used for tagging the IPI PAN Corpus (Dębowski 2003, 2004), as well as in two taggers developed by Maciej Piasecki's M.Sc. students at the Technical University of Wrocław (Krupa 2003, Gaweł 2004, Piasecki and Gaweł 2005).

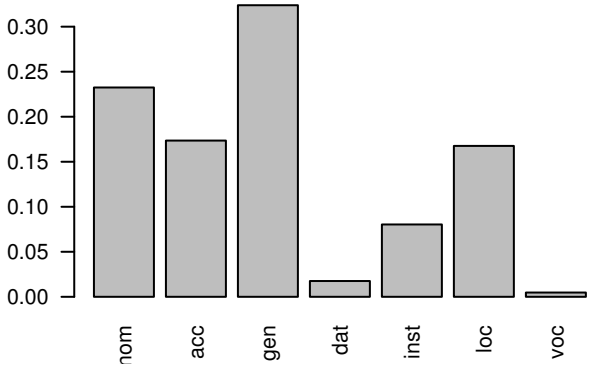

Figure 6: Frequencies of grammatical cases in `sample`.

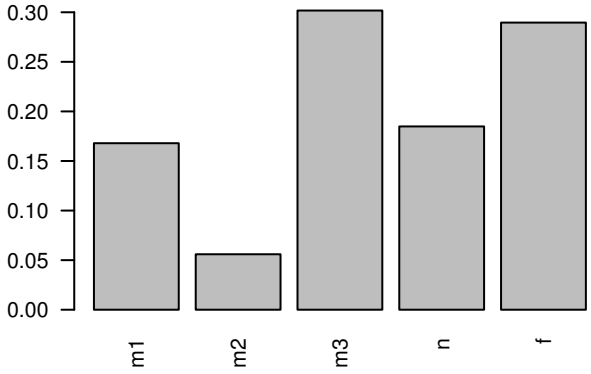

Figure 7: Frequencies of grammatical genders in `frek`.

Although the tagging techniques used at IPI PAN and at the Wrocław University of Technology are quite different, they share the general paradigm within which a certain (statistical in Dębowski 2003, 2004 and genetic in Krupa 2003, Gaweł 2004, Piasecki and Gaweł 2005) model of possible sequences of tags in a language is constructed, whose parameters are learned from a manually tagged corpus. For example, such a model may learn that a noun is very likely while a verb is very unlikely after a preposition. When this model is subsequently used in the disambiguation mode, it may decide that a post-prepositional word with a verbal interpretation and a nominal interpretation is actually a noun.

A manually disambiguated corpus may also be used for the development of taggers of a different kind, e.g., rule-based taggers, where ambiguous words are disambiguated by the application of rules manually constructed by linguists. Applications of the IPI PAN Corpus to the construction of

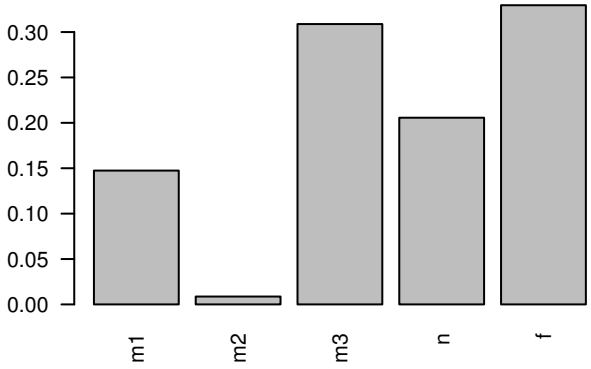

Figure 8: Frequencies of grammatical genders in `sample`.

such rule-based taggers is anticipated.

4.2. Identification of bound words

Bound words, also known as *cranberry words*, are words which only occur in a constant lexical context. There seems to be a growing theoretical interest in the description of bound words (cf., e.g., Richter and Sailer 2003) and even specialized computer collections of bound words for a given language (Sailer and Trawiński 2005). Soehn and Sailer 2003 and Sailer and Trawiński 2005 mention that there are three basic types of constructions hosting bound words (BWs):

- constructions where a BW selects a specific lexeme, as in the German *Angst* ***einjagen*** 'to scare someone', where the BW *einjagen* selects the lexeme *Angst*,
- constructions where a specific lexeme selects a BW, as in the German ***Tacheles*** *reden* 'to talk straight', where *Tacheles* is selected by *reden*,
- and constructions involving a specific lexeme selecting another specific lexeme which in turn selects a BW, as in the German *zu* ***Potte*** *kommen* 'to manage a task', where *kommen* selects *zu* and the BW *Potte*.

An initial experiment in the automatic identification of Polish bound words was performed by the author, aiming at identifying those BWs which are adjacent to a specific lexeme. The experiment did not apply any inferential statistics (like t test or χ square), but it applied various artificially set cutoff points; in particular, only lexemes whose forms occurred 20 times or more were checked for boundness. Altogether four subexperiments were

performed, for `sample` and for `wstepny`, with the aim of identifying either post-lexical or pre-lexical BWs.

The best 20 results of this experiment, as performed on the `sample` corpus for post-lexical BWs, after removing proper names (e.g., *Los* ***Angeles***) are: *z* ***dala***, *po* ***cichu***, *z* ***późn***, *w* ***przeddzień***, *zamek* ***żupny***, *na* ***przekór***, *komórka* ***tuczna***, *od* ***niechcenia***, *po* ***trosze***, *nie* ***omieszkać***, *po* ***prostu***, *collegium* ***medicum***, *ex* ***aequo***, *w* ***zupełności***, *na* ***odchodne***, *o* ***nieagresji***, *biesiadny* ***savoir***, *za* ***młodu***, *lekka* ***atletyka***, *w* ***dyrdy***. The majority of these seem to be clear cases of BWs, with a few near-BWs, such as *po* ***cichu*** (cf. *z* ***cicha***) *collegium* ***medicum*** (cf., e.g., *Forum* ***Medicum***) and *lekka* ***atletyka*** (cf., e.g., ***atletyka*** *terenowa*), and with some non-BWs, e.g., *z* ***późn***, which stems from legal texts containing the sequence *z późn. zm.* 'with later modifications', *zamek* ***żupny***, *komórka* ***tuczna*** and *biesiadny* ***savoir***. The application of statistical tests is expected to result in greater precision and recall.

4.3. Valence acquisition

A more advanced potential use of morphosyntactically annotated corpora is for the automatic acquisition of so-called valence (or subcategorisation) information, i.e., information about the arguments of predicates (usually verbs). Valence dictionaries are crucial resources for the operation of syntactic parsers, and yet, for many languages such resources are unavailable or they are available in paper form only. In case of Polish, there is only one large valence dictionary, namely, Polański 1992, although valence information is present also in a recent general dictionary of Polish (Bańko 2000). Both dictionaries are available only in the traditional paper form. Additionally, there are two much smaller electronic valence dictionaries, created by Marek Świdziński and Zygmunt Vetulani.

Initial methods of extracting valence information from corpora were proposed in early 1990ies (Brent 1993, Manning 1993), and almost all work since then was concerned with English and German, with — to the best of our knowledge — no previous attempts at acquiring subcategorisation information for Polish. At the time of writing this article, a new project financed by the State Committee for Scientific Research (project number 3 T11C 003 28) is taking off at the Institute of Computer Science, Polish Academy of Sciences, whose aim is to develop algorithms of valence extraction suitable for languages with rich morphology and relatively free word order, such as Polish.

The very first attempts at applying to the IPI PAN Corpus some of the techniques commonly used for valence extraction from English texts are described in Przepiórkowski and Fast 2005 and Fast and Przepiórkowski

2005. The experiments consist of two steps. First, a simple shallow grammar is applied to the XML corpus sources, resulting in the identification of some nominal phrases (NPs), prepositional phrases (PPs) and verbs. Then, co-occurrences of verbs and observed valence frames (i.e., multisets of the identified NPs and PPs) are counted and various known statistics are applied to these counts, including t test, likelihood ratio, binomial hypothesis test, and a number of simpler statistics, in order to filter out those observations which resulted from errors at various stages of corpus processing (morphological analysis, disambiguation, shallow parsing).

The evaluation of the results of these preliminary experiments is encouraging. In fact, the best statistic found by these experiments, i.e., the binomial hypothesis testing of Brent 1993, when applied to high frequency observed frames, gives results which agree with the information found in manually created valence dictionaries almost to the same extent that two such manual valence dictionaries agree between themselves. The perplexed reader is referred to the two articles cited above for details.

4.4. Construction of a wordnet

After describing some actual uses of the IPI PAN Corpus which go beyond treating a corpus as a bag of words, we finish this section by presenting an intended but still only potential use of the IPI PAN Corpus, namely, for the discovery of such semantic relations between lexemes as hyponymy/hyperonymy, synonymy/antonymy, holonymy/mereonymy, etc.

WordNet, the first electronic dictionary containing such information, was created for English in late 1980ies by a team of lexicographers at Princeton, led by George Miller (Miller et al. 1990), and has been maintained and further developed ever since (cf., e.g., Fellbaum 1998 and `http://www.cogsci.princeton.edu/~wn/`). Similar dictionaries, or wordnets, have also been created for many other European languages, many of them within the projects EuroWordNet and BalkaNet. Unfortunately, Polish is one of the languages with no wordnet-like dictionary available.

While the wordnets mentioned above were developed either fully manually, or with the help of electronic mono- and bilingual dictionaries, there have been attempts at the extraction of semantic lexical relations from large corpora. For example, Hearst 1992 describes a method of finding hyponym-hyperonym pairs in a corpus by finding syntactic environments known to reflect such hyponymy relations, as in *X, Y and other Zs* (e.g., *guitar, piano, and other musical instruments*) or *such Zs as X or/and Y*, which imply that X and Y are hyponyms of Z. A Polish example, attested in the `sample` subcorpus, is: *Pacjenci... źle znoszą jabłka, gruszki, śliwki i inne owoce...* 'Patients... do not take well apples, pears, plums and other fruits...'.

While this method has very good precision (few errors are made), it has very poor recall (few hyponym/hyperonym pairs are identified), so — despite consequent improvements (Hearst 1998, Caraballo 2001) — automatic extraction of hierarchical lexico-semantic relations remains a scientific challenge. We hope that the IPI PAN Corpus will be useful in facing this challenge.

5. Conclusion

The aim of the IPI PAN Corpus project mentioned in §1 was to build a large corpus of Polish suitable not only for traditional lexicographic applications or other purely linguistic research,[8] but useful primarily in natural language processing. As described in Przepiórkowski 2004, the means to achieve this end included a well-designed, rich and properly documented tagset and the employment of various open standards for corpus encoding, such as UTF-8 and the XML Corpus Encoding Standard.

This article attempted to show that this aim has largely been achieved: the IPI PAN Corpus has already, within a year of the completion of its preliminary version, been useful in learning various facts about Polish and in various NLP applications, and it has given rise to an interesting valence acquisition project.

Let us conclude by mentioning a different good use the corpus was put to, namely, in education. First of all, the corpus has been and continues to be used at the Linguistic Engineering and Natural Language Processing lectures taught by the author at the Institute of Informatics, University of Warsaw, and at the Faculty of Mathematics and Information Science, Warsaw University of Technology, respectively. In both courses the IPI PAN Corpus provides the material for various experiments in applying statistical NLP techniques and for term projects. The corpus has also been used during a number of courses at the Institute of Polish and the Institute of English, University of Warsaw. There, the students used the Poliqarp search engine (Przepiórkowski et al. 2004, Przepiórkowski 2004), also developed within the IPI PAN Corpus project, and learned its expressive query language. The IPI PAN Corpus has also been used at these linguistic departments in teaching about the morphosyntactic system of Polish and in contrasting prescriptive rules with the actual language use.

It is our hope that the IPI PAN Corpus will continue to grow and be useful to the linguistic and computational linguistic community.

[8] See, e.g., Saloni and Wołosz 2005 and Meyer 2005 for examples of the employment of the IPI PAN Corpus in, respectively, lexicographic and syntactic research.

REFERENCES

Bańko, M. (ed.) 2000. *Inny słownik języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Bień, J. S. 1991. *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

———. 2004. "An approach to computational morphology." Kłopotek et al. (2004). 191–199.

Brent, M. R. 1993. "From grammar to lexicon: Unsupervised learning of lexical syntax." *Computational Linguistics* 19, 2. 243–262.

Caraballo, S. A. 2001. *Automatic construction of a hypernym-labeled noun hierarchy from text*. Ph. D. dissertation, Brown University.

Czerepowicka, M. and Z. Saloni. 2004. "Co skreślano i co dopisywano w korpusie *Słownika frekwencyjnego polszczyzny współczesnej*." In I. Kamińska-Szmaj (ed.), *Od starożytności do współczesności. Język–literatura–kultura. Księga poświęcona pamięci profesora Jerzego Woronczaka*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego. 381–391.

Dębowski, Ł. 2003. "A reconfigurable stochastic tagger for languages with complex tag structure." In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

———. 2004. "Trigram morphosyntactic tagger for Polish." Kłopotek et al. (2004). 409–413.

Fast, J. and A. Przepiórkowski. 2005. Automatic extraction of Polish verb subcategorization: An evaluation of common statistics. Manuscript.

Fellbaum, C. (ed.) 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gaweł, B. 2004. "Zastosowanie metod programowania genetycznego do oznaczania wyrazów w polskim tekście." Master's thesis, Wrocław University of Technology.

Hearst, M. 1992. "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of the Fourteenth International conference on computational linguistics*. Nantes, France.

———. 1998. "Automated discovery of WordNet relations." Fellbaum (1998). 131–151.

Ide, N., P. Bonhomme, and L. Romary. 2000. "XCES: An XML-based standard for linguistic corpora." In *Proceedings of the linguistic resources and evaluation conference*. Athens, Greece.

Krupa, A. 2003. "Zastosowanie metod ewolucyjnych w automatycznym oznaczaniu wyrazów w polskim tekście." Master's thesis, Wrocław University of Technology.

Kurcz, I., A. Lewicki, J. Sambor, K. Szafran, and J. Woronczak. 1990. *Słownik frekwencyjny polszczyzny współczesnej*. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Kurcz, I., A. Lewicki, J. Sambor, and J. Woronczak. 1974. Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Manuscript, University of Warsaw.

Kłopotek, M. A., S. T. Wierzchoń, and K. Trojanowski (eds.) 2004. *Intelligent information processing and web mining*. Berlin: Springer-Verlag.

——— (eds.) 2005. *Intelligent information processing and web mining*. Berlin: Springer-Verlag.

Manning, C. D. 1993. "Automatic acquisition of a large subcategorization dictionary from corpora." In *Proceedings of the 31st ACL*. 235–242.

Meyer, R. 2005. VP-fronting in Czech and Polish—a case study in corpus-oriented grammar research. Manuscript, University of Regensburg.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross, and M. K. J. 1990. "Introduction to WordNet: An online lexical database." *International Journal of Lexicography* 3, 4. 235–244.

Ogrodniczuk, M. 2003. "Nowa edycja wzbogaconego korpusu słownika frekwencyjnego." In S. Gajda (ed.), *Językoznawstwo w polsce. Stan i perspektywy*. Opole: Komitet Językoznawstwa, Polska Akademia Nauk and Instytut Filologii Polskiej, Uniwersytet Opolski. 181–190. `http://www.mimuw.edu.pl/~jsbien/MO/JwP03/`

Piasecki, M. and B. Gaweł. 2005. "A rule-based tagger for Polish based on genetic algorithm." Kłopotek et al. (2005). 247–258.

Piskorski, J., P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and M. Woliński. 2004. "Information extraction for Polish using the SProUT platform." Kłopotek et al. (2004). 227–236.

Polański, K. (ed.) 1980–1992. *Słownik syntaktyczno-generatywny czasowników polskich*. Wrocław / Kraków: Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN.

Przepiórkowski, A. 2003. "Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN." *Polonica* XXII–XXIII. 57–76.

———. 2004. *The IPI PAN corpus: Preliminary version*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Przepiórkowski, A. and J. Fast. 2005. "Baseline experiments in the extraction of Polish valence frames." Kłopotek et al. (2005). 511–520.

Przepiórkowski, A., Z. Krynicki, Ł. Dębowski, M. Woliński, D. Janus, and P. Bański. 2004. "A search tool for corpora with positional tagsets and ambiguities." In *Proceedings of the fourth international Conference on Language Resources and Evaluation, LREC 2004*. Lisbon. 1235–1238.

Przepiórkowski, A. and M. Woliński. 2003. "The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish." In *Proceedings of the* 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, EACL 2003*. 109–116.

R Development Core Team. 2004. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing

Richter, F. and M. Sailer. 2003. "Cranberry words in formal grammar." In C. Beyssade et al. (eds.), *Empirical issues in formal syntax and semantics 4*, volume 4. Presses de l'Université de Paris-Sorbonne, Paris. 155–171.

Sailer, M. and B. Trawiński. 2005. "Die Sammlung unikaler Wörter des Deutschen. Aufbauprinzipien und erste Auswertungsergebnisse." In *Proceedings of europhras 2004*. Forthcoming.

Saloni, Z. and R. Wołosz. 2005. Kure(w)stwo. Manuscript, submitted to *Poradnik Językowy*.

Soehn, J.-P. and M. Sailer. 2003. "At first blush on tenterhooks: About selectional restrictions imposed by nonheads." In G. P. Gerhard Jäger, Paola Monachesi and S. Wintner (eds.), *Proceedings of formal grammar 2003*. 149–161.

Tokarski, J. 1993. *Schematyczny indeks* a tergo *polskich form wyrazowych*. Warsaw: Wydawnictwo Naukowe PWN. Elaborated and edited by Zygmunt Saloni

Woliński, M. 2003. "System znaczników morfosyntaktycznych w korpusie IPI PAN." *Polonica* XXII–XXIII. 39–55.